

Know Thy Limits: Calibrated Uncertainty for Safer Rehabilitation

Joel Strickland

Intellegens, The Studio, Chesterton Mill, Cambridge, UK

Article Info

Article Notes

Received: February 24, 2026

Accepted: March 23, 2026

*Correspondence:

*Dr. Joel Strickland. Intellegens, The Studio, Chesterton Mill, Cambridge, UK; Email: joel@intellegens.com

©2026 Strickland J. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License.

Key Words:

Rehabilitation
Uncertainty Quantification
Wearable Monitoring
Heat Stress
Core Body Temperature
Clinical Decision-making
Conformal Prediction
Return-to-work
Risk stratification

Abstract

Rehabilitation clinicians make threshold-based decisions—exercise progression, return-to-work clearance, heat safety management—that depend on reliable physiological monitoring. Yet current noninvasive methods for estimating core body temperature provide only point predictions with no indication of reliability. A prediction of 38.3°C tells clinicians nothing about whether the true value is 38.2°C or 39.0°C, making it impossible to distinguish a trustworthy estimate from a dangerously uncertain one. This limitation is especially concerning in rehabilitation, where autonomic dysfunction, cardiovascular medications, and atypical physiology cause prediction errors to vary unpredictably across patients and conditions. This mini-review argues that calibrated uncertainty—not accuracy alone—should be a foundational requirement for physiological monitoring in rehabilitation. It introduces conformal prediction, a framework that produces statistically valid prediction intervals: when configured for 95% confidence, the true temperature falls within the stated bounds approximately 95% of the time. Proof-of-concept evidence spanning over 140,000 measurements across six high-heat operational domains demonstrates that such calibration is technically achievable in real-world settings. For rehabilitation practice, uncertainty-aware monitoring enables risk-stratified exercise progression, defensible return-to-work decisions grounded in explicit confidence bounds, and scalable telemonitoring with transparent escalation pathways. The central principle is that uncertainty should be treated as a first-class clinical signal—used to pace progression, trigger conservative action, and distinguish high-risk physiology from low-trust measurement.

Introduction

Rehabilitation clinicians routinely make threshold-based decisions—when to advance exercise intensity, whether to clear a patient for return-to-work, how to manage heat exposure during outdoor therapy—that depend on physiological monitoring. Yet the noninvasive tools available for core body temperature estimation provide only point predictions with no measure of reliability^{1,2}. A predicted temperature of 38.3°C offers no indication of whether the true value is 38.2°C or 39.0°C, leaving clinicians unable to distinguish a trustworthy estimate from a dangerously unreliable one. This invited mini-review, drawing on the author's recent work applying conformal prediction to core temperature estimation³ and the broader uncertainty quantification literature, argues that calibrated uncertainty—not accuracy alone—should be a foundational requirement for physiological monitoring in rehabilitation therapy.

Heat exposure is an increasing occupational and public health concern, amplified by climate change and more frequent extreme heat events⁴⁻⁶. In the United States, an average of 702 heat-related deaths per year were reported between 2004 and 2018, with occupational settings contributing an average of 34 deaths annually between 1992 and 2022^{7,8}. Workers in construction, firefighting,

military operations, mining, and warehousing are especially vulnerable due to hot environmental conditions, physically demanding tasks, and requirements to wear heavy protective equipment⁹. These concerns overlap directly with rehabilitation therapy, where clinicians supervise graded exertion, functional testing, work hardening, and return-to-work programs—sometimes in warm environments or among individuals with reduced thermoregulatory reserve.

The clinical consequences of elevated core body temperature are well-characterized. Minor deviations (+0.5–1.0°C above baseline) can degrade fine motor skills and cognitive tasks^{10,11}. Moderate elevations (+1.5–2.0°C) are associated with heat exhaustion^{12,13}. Hyperthermic states above 40°C can result in heat stroke and multi-organ failure^{14,13}. Importantly, functional safety limits can be reached before catastrophic heat illness manifests, and patients may not reliably perceive early physiological danger—especially those with neurological impairment or medications that mask warning symptoms¹⁴.

This review first examines why point predictions are fundamentally inadequate for threshold-based rehabilitation decisions, then introduces conformal prediction as a practical uncertainty framework with formal coverage guarantees. It reviews multi-domain feasibility evidence and translates these concepts into rehabilitation-specific applications including risk-stratified exercise progression, defensible return-to-work decisions, and scalable telemonitoring with transparent escalation pathways.

The Problem with Point Predictions

Direct measurement of core body temperature typically relies on invasive or semi-invasive sensors that are impractical for routine rehabilitation use¹. Noninvasive alternatives infer core temperature from physiological proxies such as heart rate, skin temperature, activity, and ambient conditions^{2,1}. However, these proxies carry context-dependent error that varies with hydration, acclimatization, clothing insulation, sensor placement, and transient workload changes¹. In rehabilitation populations, additional heterogeneity is common: autonomic dysfunction, cardiometabolic disease, pain-limited

movement, and medications that modify cardiovascular response can all shift the relationship between proxy signals and internal temperature.

The fundamental problem is not that these methods are inaccurate—it is that they provide no information about *when* they are inaccurate. Consider a widely used algorithm that achieves 0.3°C average error. This statistic tells us nothing about individual predictions: is this particular estimate reliable to ±0.1°C, or could it be off by 0.8°C? The error distribution is rarely uniform. Models typically perform worse at temperature extremes, during rapid physiological transitions, in unfamiliar contexts, or for individuals with atypical physiology—precisely the conditions where accurate prediction matters most for safety.

This creates an epistemological trap: where we most need to trust the model is often where we should trust it least. A rehabilitation patient with autonomic dysfunction, on beta-blockers, performing intermittent activity in a warm outdoor environment may fall outside every training distribution the model has seen. The point prediction provides no signal that this is the case (Table 1).

Why Calibrated Uncertainty Matters in Rehabilitation

Figure 1 presents an overview of an uncertainty-aware monitoring system that addresses this challenge. The argument for calibrated uncertainty rests on a simple observation: you cannot make a threshold-based decision without knowing how much to trust the prediction.

Rehabilitation decisions are frequently threshold-driven and risk-asymmetric—it may be preferable to pause a session conservatively than to miss a high-risk heat event, especially for medically complex individuals or during return-to-work trials in warm conditions¹⁵.

Consider a predicted temperature of 38.3°C, just below a 38.5°C pause threshold. With a point prediction alone, the clinician has no basis for deciding whether to continue or pause. If the prediction is reliable to ±0.1°C, continuing is reasonable. If the prediction could be off by 0.5°C, the true temperature might already exceed 38.5°C—or might be safely below 38.0°C. The point prediction provides

Table 1. Accuracy versus trustworthiness in physiological monitoring. Average error describes typical model performance; calibration describes whether the system communicates *when* predictions are unreliable. For clinical decision-making, a well-calibrated system may be more useful than an uncalibrated system with lower average error, because clinicians can adjust safety margins appropriately.

System Type	Example	Calibrated?	Clinical Implication
Point prediction only	Kalman filter, regression	No	Cannot support threshold decisions—no way to know when estimate is unreliable
Uncalibrated uncertainty	Bayesian, ensemble, MC dropout	Poor	Intervals may be systematically too narrow or wide; false confidence
Calibrated uncertainty	Conformal prediction	Yes	Intervals have interpretable coverage; enables risk-aware decisions

Note: Calibration means stated confidence intervals achieve nominal coverage (e.g., 95% intervals contain true value ≈95% of the time). This property is independent of average accuracy. Uncalibrated methods can improve uncertainty estimation but usually require careful tuning during or after model training, ideally on held-out data.

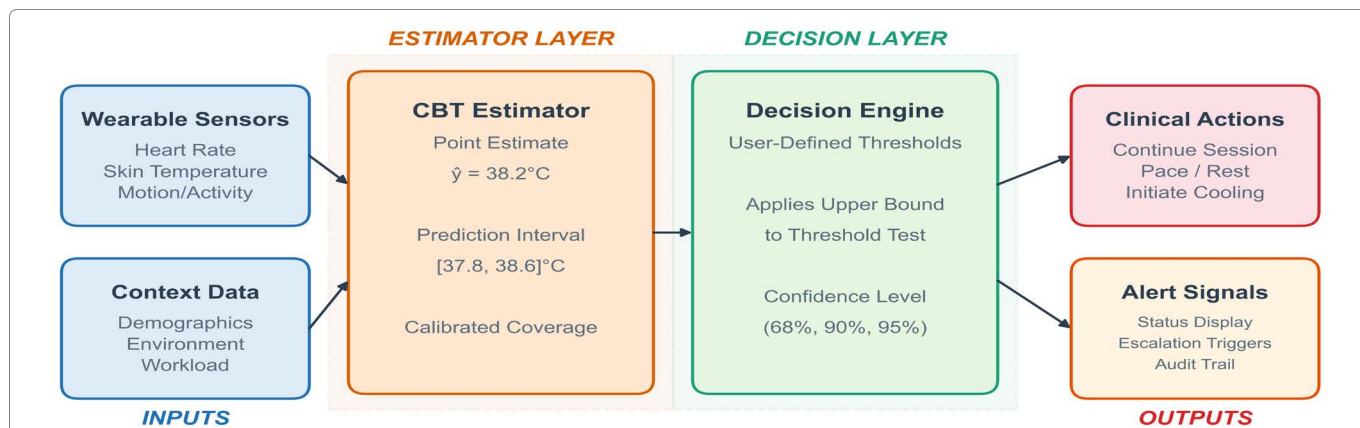


Figure 1. Uncertainty-aware core temperature monitoring workflow for rehabilitation. Wearable sensors and contextual information feed a prediction model that outputs both a point estimate and a calibrated prediction interval. Crucially, the interval width conveys how much to trust the estimate. A separate decision policy layer converts these outputs into clinical actions based on user-defined thresholds and confidence levels, enabling transparent, auditable protocols. The key insight is that even a less accurate model becomes clinically useful if its uncertainty is well-calibrated.

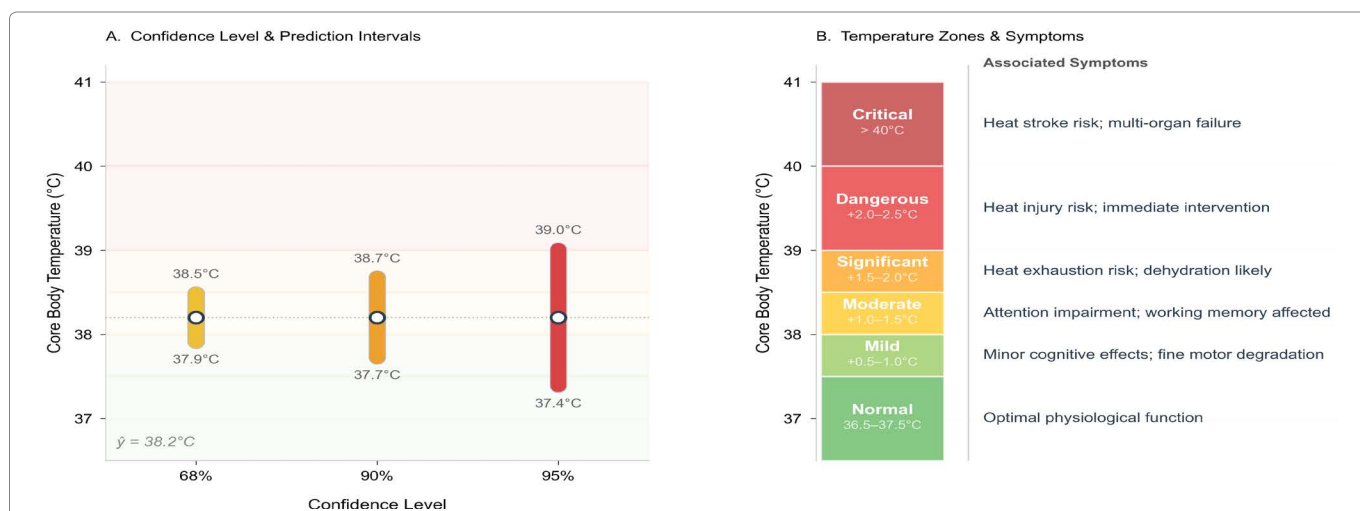


Figure 2. Calibrated uncertainty for clinical decision support. (A) Effect of confidence level on prediction intervals and clinical thresholds. The same prediction (38.2°C) is shown with intervals at 68%, 90%, and 95% confidence. Higher confidence levels produce wider intervals that may cross action thresholds—this is a feature, not a bug, as it forces conservative decisions when risk is uncertain. (B) Clinical temperature zones showing symptoms that may occur at each level—note that symptom expression is individual and context-dependent, and progression through zones is not deterministic.

no information to distinguish these scenarios. Figure 2 illustrates how different confidence levels affect threshold-based decisions.

Uncertainty-aware monitoring supports three practical functions:

1. **Risk-aware progression:** When uncertainty is high (wide intervals), clinicians can slow progression, increase rest, or repeat measurement before advancing intensity.
2. **Actionable escalation:** Upper confidence bounds can trigger conservative interventions even if the point estimate is below threshold.
3. **Auditability:** Calibrated intervals provide a

defensible rationale for protocols, enabling documentation that states intervention occurs when the 95% upper bound exceeds a specified value.

Worked Example: A patient undergoing work hardening shows a predicted core temperature of 38.3°C with a 95% interval of [37.6, 39.0°C]. The interval width (1.4°C) exceeds the quality threshold and the upper bound reaches 39.0°C—triggering a pause. After sensor adjustment and 5 minutes of rest, the prediction updates to 37.8°C with interval [37.5, 38.1°C]. The narrower interval (0.6°C) and upper bound well below threshold now support resuming activity. The difference is not just the point estimate—it is the *actionable information* provided by the interval width.

Example Threshold Protocol

- If 95% upper bound $\geq 38.5^{\circ}\text{C}$:
 - Pause, initiate cooling, reassess in 5 min
- If interval width $> 0.8^{\circ}\text{C}^*$:
 - Check sensors, fit, corroborate with symptoms
- If upper bound $\geq 39.0^{\circ}\text{C}$ for >10 min:
 - Escalate care, consider emergency protocols

*Example quality control (QC) threshold; tune to device and setting.

Conformal Prediction: A Practical Framework

For rehabilitation clinicians, the key question about any monitoring system is whether its reported confidence can be trusted—and if so, how much. Several frameworks exist for prediction-level uncertainty: Bayesian neural networks, ensemble methods (including Monte Carlo dropout), and post-hoc calibration techniques^{3,16,17}. These methods can improve uncertainty estimation, but none provides distribution-free coverage guarantees, and all require careful tuning to avoid systematically overconfident or underconfident intervals (Table 1).

Conformal prediction offers a complementary approach with distinct advantages for safety-critical applications^{18,16}. In practical terms, conformal prediction can make verifiable promises about its reliability: if configured for 95% confidence, the true temperature will fall within the stated bounds approximately 95 out of 100 times. The mechanism is straightforward. Rather than requiring strong distributional assumptions, conformal methods use a held-out calibration set to determine the distribution of prediction errors, then produce bounds around each new prediction at a selected confidence level. Under an exchangeability assumption between calibration and deployment data, the resulting intervals achieve the target coverage rate with formal statistical guarantees.

For rehabilitation, this makes uncertainty a controllable parameter. Confidence can be set more conservatively (e.g., 99%) for high-risk patients or high-heat environments, producing wider intervals that trigger earlier intervention. Conversely, when interruptions are costly and risk is low, narrower intervals at 80% or 90% confidence may be appropriate. The confidence level becomes a clinical dial that can be adjusted based on context, rather than a hidden model assumption.

A further refinement is stratified conformal calibration, which addresses the fact that prediction error is not uniform across conditions³. Error typically varies across temperature ranges—often larger at elevated temperatures where monitoring is most critical. Stratifying calibration by predicted temperature bands allows bounds to widen where the model tends to be less accurate and tighten where it is more reliable. For a rehabilitation clinician, this means the system automatically becomes more cautious as temperatures approach dangerous levels—precisely when conservative

action matters most—while avoiding unnecessary interruptions when temperatures are clearly safe.

Feasibility Evidence: Calibrated Uncertainty in Practice

Is calibrated uncertainty achievable in real-world physiological monitoring? Recent work provides proof-of-concept evidence³. The study employed a hybrid deep learning architecture—bidirectional Long Short-Term Memory (LSTM) networks combined with dense layers—developed and evaluated using over 140,000 physiological measurements from 251 participants across six high-heat operational domains: wildland firefighting, motorsport, mining, nuclear plant work, explosive ordnance disposal, and factory settings.

Model inputs included heart rate, skin temperature, ambient temperature, activity level, and demographic variables (age, sex, body mass, height, body surface area), all commonly available from wearable devices. Calibrated prediction intervals were generated using inductive conformal prediction with stratified calibration across a held-out set of approximately 24,000 measurements; full methodological detail is provided in³. The central finding was clinical, not statistical: the system could reliably distinguish trustworthy predictions from uncertain ones. Average accuracy was comparable to existing methods (0.29°C vs 0.34°C for a benchmark algorithm²), but the system's stated confidence bounds matched reality 12 times more faithfully—meaning its uncertainty estimates could actually be trusted for threshold-based decisions.

This distinction matters. When the system reported 95% confidence bounds, the true temperature fell within those bounds approximately 95% of the time—the algorithm “knew when it didn't know.” Stratified calibration adapted interval width to local error behavior: bounds widened at elevated temperatures and narrowed where the model was more reliable. For a clinician supervising graded exercise or return-to-work trials, the monitoring system communicates not just *what* the temperature is, but *how much* to trust it—enabling informed rather than blind threshold decisions.

While no other studies have applied conformal prediction specifically to core temperature estimation in rehabilitation, uncertainty quantification is an active area in related physiological monitoring domains. Bayesian and ensemble approaches have been explored for blood glucose prediction¹⁹, and conformal prediction has recently been applied to wearable cuffless blood pressure estimation with calibrated confidence intervals²⁰. The application of distribution-free uncertainty quantification to thermoregulatory monitoring remains an emerging field, and the evidence reviewed here represents an early proof-of-concept that the approach is technically viable.

The translational lesson is not that clinics must adopt any particular algorithm, but that uncertainty calibration is technically feasible and should be demanded from monitoring systems. The principle is to separate estimation from decision policy (Figure 1): the estimator provides a prediction and calibrated interval; the decision policy determines actions based on patient risk and program goals.

Rehabilitation-Specific Applications

Once calibrated uncertainty is available, rehabilitation applications are straightforward. In **graded exercise therapy**, uncertainty-aware monitoring can identify periods of stable low-risk physiology versus periods requiring conservative adjustment^{21,22}. When intervals are narrow and temperature stable, progression can proceed; when intervals widen or upper bounds approach thresholds, intensity can be reduced or recovery extended. This is particularly relevant for cardiopulmonary rehabilitation, outdoor therapy programs, and sessions involving patients with uncertain thermoregulatory reserve.

Critically, different rehabilitation populations present distinct thermoregulatory profiles that affect both prediction accuracy and clinical risk. Patients with **neurological conditions**—spinal cord injury, stroke, multiple sclerosis—may have impaired autonomic thermoregulation, reduced sweating capacity below lesion level, and vasomotor dysfunction that disrupts normal heat dissipation. **Cardiovascular** populations face compounding factors: beta-blockers blunt heart rate response (a key model input), reduced cardiac output limits convective heat transfer, and diuretics alter hydration status. **Musculoskeletal** rehabilitation patients generally retain intact thermoregulation but may have reduced activity tolerance or use protective equipment that traps heat. This heterogeneity is precisely why calibrated uncertainty matters: the same algorithm will behave differently across these groups, and prediction intervals will naturally widen for populations whose physiology diverges from training data—signaling the need for conservative thresholds rather than silent overconfidence.

In **work hardening and return-to-work programs**, monitoring can provide defensible documentation of heat tolerance during simulated tasks^{21,22}. Calibrated intervals support accommodation decisions—work-rest cycles, cooling access, hydration protocols, and protective equipment modifications can all be tied to objective physiological data with explicit confidence bounds.

For **telerehabilitation and remote monitoring**, uncertainty-aware triage can support scalable escalation pathways²³:

- Self-management prompts when upper bounds approach a conservative threshold

- Clinician messaging when sustained upper bounds exceed a higher threshold
- Urgent escalation when bounds exceed critical levels or symptoms co-occur

Implementation considerations matter significantly. Sensor placement, contact quality, and rehabilitation-specific equipment (braces, compression garments, assistive devices) all influence signal quality¹. Uncertainty should therefore be treated as a quality signal: a widening interval can prompt sensor troubleshooting rather than escalation, helping clinicians distinguish “high-risk physiology” from “low-trust measurement.”

Equity and alarm fatigue deserve attention. Rehabilitation populations are heterogeneous, and wearable algorithms may behave differently across ages, comorbidity profiles, and medication regimens. Calibrated uncertainty reduces the harm of silent failure by making “unknown” states visible—if the model is less reliable in a subgroup, intervals widen, signaling the need for conservative action. Conservative thresholds will increase interruptions, but the solution is to *tune conservatism to context*: high-risk patients warrant 95% or 99% bounds; lower-risk patients may tolerate 80% or 90% bounds. Uncertainty-aware systems *make these trade-offs explicit and adjustable*.

Limitations and Future Directions

All uncertainty quantification methods rely on assumptions that may not hold in deployment. Conformal prediction requires exchangeability between calibration and deployment data; Bayesian methods require well-specified priors; ensemble methods require diverse model components. If calibration data do not reflect rehabilitation populations—different age distributions, comorbidity profiles, medication use—interval reliability can degrade²⁴. This concern is especially acute given the thermoregulatory heterogeneity across rehabilitation subgroups. Neurological populations (spinal cord injury, stroke) with impaired autonomic regulation, cardiovascular patients on medications that alter physiological signals, and pulmonary rehabilitation patients with ventilatory constraints on heat dissipation each present distinct calibration challenges. Because conformal prediction’s coverage guarantees depend on exchangeability between calibration and deployment data, population-specific calibration sets—built from rehabilitation cohorts rather than occupational datasets alone—may be necessary to maintain interval reliability²⁴. Developing and validating such cohorts across neurological, cardiovascular, and pulmonary rehabilitation populations should be a priority for translational research.

Prospective studies should evaluate not only predictive metrics but clinically meaningful outcomes: symptom burden, adverse events, session completion rates, patient

acceptability, and alarm fatigue. Decision-threshold studies are particularly valuable. Comparing protocols that trigger actions based on point estimates versus upper confidence bounds can quantify real trade-offs between false alarms and missed heat events. Such studies would help standardize protocols and provide evidence for confident clinical adoption. The conceptual and methodological framework presented in this review—distinguishing estimation from decision policy, with calibrated uncertainty as the bridge—can inform the design of such trials.

Additional research should examine performance under conditions common in rehabilitation: intermittent sensor contact, indoor-outdoor transitions, and atypical thermoregulatory profiles including patients with autonomic dysfunction or cardiovascular limitations.

Conclusion

Calibrated uncertainty—not accuracy alone—should be a foundational requirement for physiological monitoring in rehabilitation. Exercise progression, return-to-work clearance, and telemonitoring escalation all depend on threshold decisions that are unsafe without reliable confidence bounds. Recent evidence demonstrates that calibrated prediction intervals are technically achievable in real-world physiological monitoring. The path forward requires rehabilitation-specific validation cohorts, prospective studies comparing point-prediction versus uncertainty-aware protocols, and systematic evaluation of clinical outcomes—ensuring that uncertainty quantification becomes standard practice as wearable monitoring expands in rehabilitation.

Conflict of Interest

The author is employed by Intellegens, a company that develops machine learning software including uncertainty quantification methods. The views expressed are the author's own. Independent replication of the calibration results reported herein, particularly in rehabilitation-specific populations, will be essential to validate the clinical recommendations made in this review.

Funding

No specific funding was received for this mini-review.

References

1. Dolson CM, Harlow ER, Habeeb CM, et al. Wearable sensor technology to predict core body temperature: a systematic review. *Sensors*. 2022; 22(19): 7639.
2. Buller MJ, Tharion WJ, Chevront SN, et al. Estimation of human core temperature from sequential heart rate observations. *Physiol Meas*. 2013; 34(7): 781–798.
3. Strickland J, Ghisoni M, Marshall H, et al. Degrees of uncertainty: conformal deep learning for non-invasive core body temperature prediction in extreme environments. *Communications Engineering*. 2025; 4: 219.
4. Flouris AD, Dinas PC, Ioannou LG, et al. Workers' health and productivity under occupational heat strain: a systematic review and meta-analysis. *Lancet Planet Health*. 2018; 2(12): e521–e531.
5. Fatima SH, Rothmore P, Giles LC, Varghese BM, Bi P. Extreme heat and occupational injuries in different climate zones: a systematic review and meta-analysis. *Environ Int*. 2021; 148: 106384.
6. Patel L, Conlon KC, Sorensen C, et al. Climate change and extreme heat events: how health systems should prepare. *NEJM Catalyst Innovations in Care Delivery*. 2022; 3(7): CAT.21.0454.
7. Vaidyanathan A, Malilay J, Schramm P, Saha S. Heat-related deaths—United States, 2004–2018. *MMWR Morb Mortal Wkly Rep*. 2020; 69(24): 729–734.
8. United States Environmental Protection Agency. A closer look: heat-related workplace deaths [Internet]. 2025 [cited 2026 Feb 20]. Available from: <https://www.epa.gov/climate-indicators/understanding-connections-between-climate-change-and-human-health>.
9. Habibi P, Moradi G, Dehghan H, et al. Climate change and heat stress resilient outdoor workers: findings from systematic literature review. *BMC Public Health*. 2024; 24: 1711.
10. Tansey EA, Johnson CD. Recent advances in thermoregulation. *Adv Physiol Educ*. 2015; 39(3): 139–148.
11. Taylor L, Watkins SL, Marshall H, Dascombe BJ, Foster J. The impact of different environmental conditions on cognitive function: a focused review. *Front Physiol*. 2016; 6: 372.
12. Schmit C, Hausswirth C, Le Meur Y, Duffield R. Cognitive functioning and heat strain: performance responses and protective strategies. *Sports Med*. 2017; 47(7): 1289–1302.
13. Gauer R, Meyers BK. Heat-related illnesses. *Am Fam Physician*. 2019; 99(8): 482–489.
14. Bouchama A, Knochel JP. Heat stroke. *N Engl J Med*. 2002; 346(25): 1978–1988.
15. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription. 11th ed. Philadelphia: Wolters Kluwer; 2022.
16. Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res*. 2008; 9: 371–421.
17. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511 [preprint]. 2021 [cited 2026 Feb 20]. Available from: <https://arxiv.org/abs/2107.07511>.
18. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. *J Am Stat Assoc*. 2018; 113(523): 1094–1111.
19. Zhu T, Li K, Herrero P, Georgiou P. Deep learning for diabetes: a systematic review. *IEEE J Biomed Health Inform*. 2021; 25(7): 2744–2757.
20. Shen Z, Chakraborti T, Banerji CRS, Ding X. Conformal prediction quantifies wearable cuffless blood pressure with certainty. *Sci Rep*. 2025; 15(1): 26697.
21. Washington State Department of Labor & Industries. Work Rehabilitation Guideline. November 2021.
22. Notley SR, Flouris AD, Kenny GP. Physiological monitoring for occupational heat stress management: recent advancements and remaining challenges. *Appl Physiol Nutr Metab*. 2025; 50: 1–14.
23. Lee AC, Davenport TE, Randall K. Telerehabilitation in physical therapist practice: a clinical practice guideline from the American Physical Therapy Association. *Phys Ther*. 2024; 104(4): pzae015.
24. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. Conformal prediction beyond exchangeability. *Ann Stat*. 2023; 51(2): 816–845.